

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA **HỆ THỐNG THÔNG TIN**



MÔN HỌC DỮ LIỆU LỚN

BÀI THỰC HÀNH LAB 2

Giảng viên: ThS. Nguyễn Hồ Duy Trí

Sinh viên thực hiện:

Ngô Thùy Yến Nhi

MSSV: 21521230

TP. HỒ CHÍ MINH, NĂM 2024

NHẬN XÉT CỦA GIÁO VIÊN

Bài 1: Giới thiệu từng bước thực hiện và minh họa bằng ảnh chụp màn hình. Cấu trúc cây thư mục

/DHQG-HCM/UIT

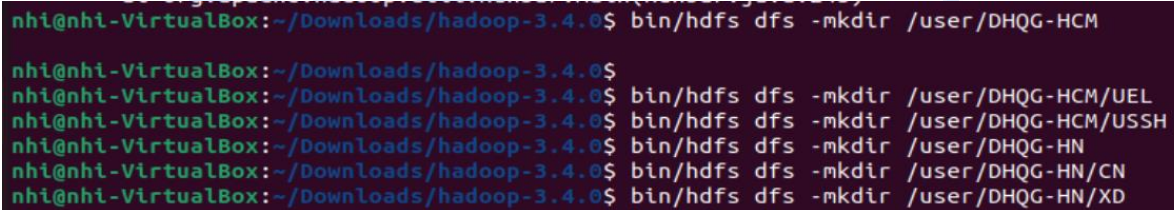
/DHQG-HCM/UEL

/DHQG-HCM/USSH

/DHQG-HN/CN

/DHQG-HN/XD

- Tạo các thư mục cần thiết



```
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/DHQG-HCM
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/DHQG-HCM/UEL
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/DHQG-HCM/USSH
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/DHQG-HN
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/DHQG-HN/CN
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hdfs dfs -mkdir /user/DHQG-HN/XD
```

- Kiểm tra thông qua trình duyệt localhost:9870

Hadoop

Overview Datanodes Datanode Volume Failures Snapshot Startup Progress

Utilities

Browse Directory

/user Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	nhi	supergroup	0 B	Oct 10 19:19	0	0 B	DHQG-HCM
drwxr-xr-x	nhi	supergroup	0 B	Oct 10 19:20	0	0 B	DHQG-HN
drwxr-xr-x	nhi	supergroup	0 B	Oct 10 10:04	0	0 B	nhi

Showing 1 to 3 of 3 entries Previous 1 Next

Bài 2: Nộp minh họa các bước làm. Download WordCount v2.0 từ [link](#)

- Thực thi combine file code và nén thành file .jar

```
VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hadoop com.sun.tools.javac.Main /home/nhi/lab2/02/WordCount2.java
VirtualBox:~/Downloads/hadoop-3.4.0$ jar cf wc.jar /home/nhi/lab2/02/WordCount*.class
```

- Tạo 2 tệp input cần thiết

```
Hello Hadoop, Goodbye to Hadoop.
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ mkdir -p ~input/WordCount
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ nano ~input/WordCount/file1.txt
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ nano ~input/WordCount/file2.txt
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ cat ~input/WordCount/file2.txt
Hello Hadoop, Goodbye to Hadoop.
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ cat ~input/WordCount/file1.txt
Hello World, Bye World!
```

- Đẩy 2 file vừa tạo lên môi trường phân tán, sau đó thực thi câu lệnh chạy và xuất ra kết quả

```

nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hdfs dfs -put ~input/WordCount/* inputs/WordCount
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hadoop jar WordCount.jar WordCount inputs/WordCount outputs/WordCount
JAR does not exist or is not a normal file: /home/nhi/Downloads/hadoop-3.4.0/WordCount.jar
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hadoop jar wc.jar WordCount inputs/WordCount outputs/WordCount
JAR does not exist or is not a normal file: /home/nhi/Downloads/hadoop-3.4.0/wc
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hadoop jar wc.jar WordCount inputs/WordCount outputs/WordCount

```

```

Total megabyte-milliseconds taken by all reduce tasks=4881408
Map-Reduce Framework
  Map input records=2
  Map output records=9
  Map output bytes=93
  Map output materialized bytes=123
  Input split bytes=246
  Combine input records=9
  Combine output records=9
  Reduce input groups=8
  Reduce shuffle bytes=123
  Reduce input records=9
  Reduce output records=8
  Spilled Records=18
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=169
  CPU time spent (ms)=2640
  Physical memory (bytes) snapshot=722866176
  Virtual memory (bytes) snapshot=8193699840
  Total committed heap usage (bytes)=479199232
  Peak Map Physical memory (bytes)=260620288
  Peak Map Virtual memory (bytes)=2726420480
  Peak Reduce Physical memory (bytes)=202309632
  Peak Reduce Virtual memory (bytes)=2745282560
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
WordCount2$TokenizerMapper$CountersEnum
  INPUT_WORDS=9
File Input Format Counters
  Bytes Read=57
File Output Format Counters
  Bytes Written=67

```

- Chạy lệnh in ra kết quả lưu trong file bắt đầu bằng “part”

```

nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hdfs dfs -cat output/WordCount/part*

```

- Kết quả thực thi

```

Bye      1
Goodbye  1
Hadoop,  1
Hello    2
World!   1
World,   1
hadoop.  1
to       1

```

Bài 3: Nộp source code và hình chụp kết quả chạy chương trình. Download dữ liệu trong đường dẫn đã cung cấp

- Chạy các lệnh tương tự, nén thành file .jar và chạy lệnh hadoop sau khi copy file input vào HDFS.

```

prob(38|3951) = 0.18095239
prob(38|3950) = 0.21052632
prob(38|3949) = 0.3
prob(38|3948) = 0.125
prob(38|3947) = 0.12820514
prob(38|3946) = 0.08791209
prob(38|3945) = 0.055555556
prob(38|3942) = 0.1509434
prob(38|3941) = 0.13333334
prob(38|3940) = 0.1764706
prob(38|3939) = 0.25714287
prob(38|3938) = 0.07692308
prob(38|3936) = 0.32258064
prob(38|3935) = 0.18518518
prob(38|3933) = 0.18181819
prob(38|3932) = 0.08064516
prob(38|3931) = 0.2
prob(38|3930) = 0.071428575
prob(38|3928) = 0.055555556
prob(38|3927) = 0.2
prob(38|3926) = 0.13636364
prob(38|3925) = 0.16666667
prob(38|3924) = 0.13636364
prob(38|3923) = 0.14782609
prob(38|3922) = 0.2413793
prob(38|3921) = 0.14285715
prob(38|3920) = 0.18181819
prob(38|3919) = 0.18181819
prob(38|3918) = 0.11111111
prob(38|3917) = 0.24444444
prob(38|3916) = 0.26666668
prob(38|3915) = 0.28
prob(38|3914) = 0.14583333
prob(38|3913) = 0.10344828
prob(38|3912) = 0.16513762
prob(38|3911) = 0.5
prob(38|3910) = 0.083333336
prob(38|3908) = 0.0625
prob(38|3907) = 0.17283951
prob(38|3906) = 0.1875
prob(38|3905) = 0.5
prob(38|3904) = 0.9765396
prob(38|3903) = 0.30864197
prob(38|3900) = 0.10869565

```

- Kết quả khi mua sản phẩm 3270 trong khi đã mua sản phẩm 13368 là

```

Peak Reduce Virtual Memory (bytes)=2749155200
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=4167490
File Output Format Counters
Bytes Written=216847079
VirtualBox: ~/Downloads/hadoop-3.4.0$ bin/hdfs dfs -cat output/test12/part* | grep *3270|13368"
b(3270|13368) = 0.07692308

```

Bài 4: Nộp source code và hình chụp kết quả chạy chương trình. Dữ liệu đã có sẵn trong thư mục data

- Thực hiện tương tự như các bài tập trên

```

nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hadoop com.sun.tools.javac.Main /home/nhi/lab2/MarketPrice/MarketPrice.java
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ jar cf MarketPrice.jar /home/nhi/lab2/MarketPrice/MarketPrice*.class
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ cat /home/nhi/Downloads/hadoop-3.4.0/input/MarketPrice/industryandtrade_market_price.csv_39.csv
cat: invalid option -- '3'
Try 'cat --help' for more information.
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ head -n 5 /home/nhi/Downloads/hadoop-3.4.0/input/MarketPrice/industryandtrade_market_price.csv_39.csv
ten,gia,donvitinh,ngaycapnhat
"Bánh chưng loại 1,5kg/cái","180000.0","Cái","2022-04-12 00:00:00"
"Bánh hộp Danisa butter cookie 454gr","140000.0","Hộp","2022-04-12 00:00:00"
"Bắp cải Đà Lạt (lặt sạch)","20000.0","Kg","2022-04-12 00:00:00"
"Bia Heineken lon (thùng 24 lon)","410000.0","Thùng","2022-04-12 00:00:00"
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ mkdir -p ~input/MarketPrice
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ head -n 5 /home/nhi/Downloads/hadoop-3.4.0/~input/MarketPrice/industryandtrade_market_price.csv_39.csv
ten,gia,donvitinh,ngaycapnhat
"Bánh chưng loại 1,5kg/cái","180000.0","Cái","2022-04-12 00:00:00"
"Bánh hộp Danisa butter cookie 454gr","140000.0","Hộp","2022-04-12 00:00:00"
"Bắp cải Đà Lạt (lặt sạch)","20000.0","Kg","2022-04-12 00:00:00"
"Bia Heineken lon (thùng 24 lon)","410000.0","Thùng","2022-04-12 00:00:00"
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hdfs dfs -mkdir inputs/MarketPrice
nhi@nhi-VirtualBox:~/Downloads/hadoop-3.4.0$ bin/hdfs dfs -put ~input/MarketPrice/* inputs/MarketPrice

```

- In ra kết quả thực thi

"Bánh chưng loại 1,5kg/cái"	Average Price: 125000.0	Min Price: 100000.0	Max Price: 150000.0
"Bông cải xanh, bông cải trắng Đà Lạt"	Average Price: 35000.0	Min Price: 35000.0	Max Price: 35000.0
"Cà điều hồng (con trên 0,7k/g)"	Average Price: 65000.0	Min Price: 65000.0	Max Price: 65000.0
"Cá lóc nuôi bè (0,5kg/con)"	Average Price: 80000.0	Min Price: 80000.0	Max Price: 80000.0
"Đồ 0,0015-V"	Average Price: 12452.353	Min Price: 12410.0	Max Price: 12590.0
"Đồ 0,055-II"	Average Price: 12175.883	Min Price: 12110.0	Max Price: 12390.0
"Gà mái ta làm sẵn (con 1-1,5kg)"	Average Price: 120000.0	Min Price: 120000.0	Max Price: 120000.0
"Gà tam hoàng làm sẵn (con 1 - 1,5kg)"	Average Price: 75000.0	Min Price: 75000.0	Max Price: 75000.0
"Nước mắm chai 0,65lít 300 Liên Thành nhãn vàng"	Average Price: 38000.0	Min Price: 38000.0	Max Price: 38000.0
"Thịt nạc (dằm, vai, đùi)"	Average Price: 190000.0	Min Price: 190000.0	Max Price: 190000.0
"Vải Tejin nội khổ 1,40m"	Average Price: 100000.0	Min Price: 100000.0	Max Price: 100000.0
"Vịt làm sẵn (con 2-2,2kg)"	Average Price: 160000.0	Min Price: 160000.0	Max Price: 160000.0
Bia Heineken lon (thùng 24 lon)	Average Price: 400000.0	Min Price: 400000.0	Max Price: 400000.0
Bia Tiger lon (thùng 24 lon)	Average Price: 340000.0	Min Price: 340000.0	Max Price: 340000.0
Bia lon 333 Sài Gòn (thùng 24 lon)	Average Price: 250000.0	Min Price: 250000.0	Max Price: 250000.0
Bánh hộp Danisa butter cookie 454gr	Average Price: 130000.0	Min Price: 130000.0	Max Price: 130000.0
Bánh hộp thiết Bibica Goody Gold 450gr	Average Price: 95000.0	Min Price: 95000.0	Max Price: 95000.0
Bí xanh	Average Price: 17000.0	Min Price: 17000.0	Max Price: 17000.0
Bưởi năm roi (da xanh) (trái 1 kg)	Average Price: 45000.0	Min Price: 45000.0	Max Price: 45000.0
Bưởi năm roi (trái 1kg)	Average Price: 40000.0	Min Price: 40000.0	Max Price: 40000.0
Bắp cải Đà Lạt	Average Price: 18000.0	Min Price: 18000.0	Max Price: 18000.0
Bắp cải Đà Lạt (lặt sạch)	Average Price: 17000.0	Min Price: 17000.0	Max Price: 17000.0
Bột ngọt Ajinomoto (VN 453gr/gói)	Average Price: 30000.0	Min Price: 30000.0	Max Price: 30000.0
Cam sành (loại 4-5 trái/kg)	Average Price: 25000.0	Min Price: 25000.0	Max Price: 25000.0
Chanh giầy	Average Price: 18000.0	Min Price: 18000.0	Max Price: 18000.0
Chân giò	Average Price: 85000.0	Min Price: 85000.0	Max Price: 85000.0
Chả bò (CS. Định)	Average Price: 220000.0	Min Price: 220000.0	Max Price: 220000.0
Chả lụa (Phượng Nam)	Average Price: 220000.0	Min Price: 220000.0	Max Price: 220000.0